**SQL Project – Movie Data**

Hello everybody, the following is a project to practice even more on my data analysis skills. Unlike my previous project on used car data this one is based on movie data from 1980 through 1991. In this project I wanted to learn about more complex SQL problems and wanted to include the use of CTEs in some solutions. Below I provided the data and what it includes.

|  |
| --- |
| **Data** |

The data was retrieved from Kaggle and includes the following variables:

A screenshot of a phone

AI-generated content may be incorrect.

You can access the data page from Kaggle using the following link -

<https://www.kaggle.com/datasets/danielgrijalvas/movies/data>

|  |
| --- |
| **SQL** |

1. Which film has the highest score with more than 300,000 votes between the year 1985 – 1990

* The first thing I decided to do is determine the variables I needed. So, I needed the year, name, company, score, and votes.
* I decided to include company so that the stakeholder has more context about the film other than the name.
* From here we separated the movies to show only those with 300,000 votes or above
* We also made sure to that the year was between the year 1985 - 1990
* Lastly, we ordered it by score and descending to display the highest scored film at the top.

Query:

A black screen with white text

AI-generated content may be incorrect.

Result:

A screenshot of a computer

AI-generated content may be incorrect.

From the result above we can determine that the highest scored film with more than 300,000 votes was tied between 2 movies Aliens and Full Metal Jacket with a score of 8.3. In second there is a three-way tie with Die Hard, My Neighbor Totoro, and Indiana Jone and the Last Crusade with a score of 8.2.

1. List all companies that have produced more than 10 movies and have an average gross over $50 million.

* First, we need to retrieve the company, the number of films, and the average profit of the studio.
* We group it by data so that it knows how to count the number of films by studio.
* We now make it where the number of movies is greater than 10 and the profit is more than $50 Million
* I decided to order it by profit to show the most profitable studio with more than 10 films

Query: A screen shot of a computer

AI-generated content may be incorrect.

Result:

A screenshot of a computer

AI-generated content may be incorrect.

From the result above I can determine that the most profitable studio with more than 10 movies is Touchstone Pictures with 25 movies and a profit of $73,101,036.48. In second is Paramount pictures with 80 films and a profit of $66,752,109.61.

1. Which film is profitable and has a score less than 5

* First, we need to retrieve the name, score, gross, budget, and profit by subtracting the gross by the budget.
* We then filter it to only show films with a score less than 5
* We then order it by profit and descending so that we get the most profitable movie on top

Query:

A black screen with white text

AI-generated content may be incorrect.

Result:

A screenshot of a computer screen

AI-generated content may be incorrect.

From the results above I can determine that Jaws 3-D was the highest profitable movie with $67,487,055 and a score of 3.7. Behind Jaws 3-D is a film Staying Alive with a profit of $42,892,670 and a score of 4.7. I wanted to find these films because there are always those films that are rated low but are profitable because of their important name, or funny stories that gravitate people to watch them no matter what their score.

1. Which country has the highest profitable film from a country other than the United States

* First, we need to retrieve the country, name, company, gross, budget, and the profit.
* We then filter out all the countries that are not the United States
* We then order it by profit descending to get the highest profitable movie at the top
* We then limit it to 1.

Query: A black screen with white text

AI-generated content may be incorrect.

Result:



From the results above we can determine that the most profitable movie not from the United States is the movie Crocodile Dundee from Australia with a profit of $319,403,506.

1. What is the least profitable R rated movie from the United States

* First, we need to retrieve the name, rating, country, company, gross, budget, and profit.
* We filter it to only show R rated movie and that the country is Unite States.
* We order it by profit ascending to show the least profitable film.
* Lastly limit it to 1 so that it shows only the top least profitable film.

Query:

A black screen with white text

AI-generated content may be incorrect.

Result:



From the results above we can determine that the movie Hudson Hawk from TriStar Pictures is the least profitable R rated movie from the United States with a loss of $47,781,920.

1. The top 20 studio with their highest profit movie

* First, we needed to retrieve the company, name, and profit.
* This is where I realized that I would need to create a CTE to get the profit from a film.
* I also created another CTE so that I can get the most profitable film from a studio.
* We then join both CTEs on the variable of company as well as profit.
* We order the results by profit descending to the most profitable movie at the top
* Lastly limit it to 20 to get the top 20 studios.

Query:

A computer screen shot of a program

AI-generated content may be incorrect.

Result:

A screenshot of a computer

AI-generated content may be incorrect.

From the results above we can determine that Universal Pictures E.T. is the most profitable studio move from Universal Pictures with a profit of $782,410,554. Behind is Star Wars: Episode 5 from Lucasfilm with a profit of $520,375,067. And, lastly in 3rd is Ghost from Paramount Pictures with a profit of $483,703,557

1. Find all movies where the budget was higher than the average budget for their genre.

* First, we need to retrieve the company, name, budget, and the average budget
* I created a CTE to retrieve the average budget from each genre.
* We join the CTE and the dataset on the condition that they equal in genre.
* We then filter the result to only show those films where the budget is higher than the average budget for the film.

Query:

A computer screen with white text

AI-generated content may be incorrect.

Result:

A screenshot of a computer

AI-generated content may be incorrect.

From the results above we can see that we got results where the budget is higher than the average budget. On the first result we can see that the film The Shining had a budget of $27,998,772 and the average for a drama film was $12,975,776. We can see that the budget for The Shining was higher than the average budget for drama films. The second result is the adventure film The Blue Lagoon that had a budget of $54,353,106 and the average budget for an adventure film is $19,274,470.

1. List movies that had a lower score than the average score of other movies released in the same year but received more than 100,000 votes.

* This is very similar to the previous question and first we need to retrieve the name, year, votes, score, and the avg score.
* Second, we create a CTE to get the average score of each year.
* We then join the dataset with the CTE on the condition that the year is equal.
* Lastly, we filter the results to only show those where the score is greater than the average score and that has more than 100,000 votes.

Query:

A screen shot of a computer program

AI-generated content may be incorrect.

Result:

A screenshot of a table

AI-generated content may be incorrect.

From the results above I can see that there are some films where the score is less than the average of the year and has more than 100,000 votes. The first one we can see is Friday the 13th with a score of 6.4 and the year’s average is 6.51 in 1980. The film as well gathered around 123,000 votes that contributed to the score of the film.

1. Highest Profitable Movie from each month

* First, we need to retrieve the month, max profit, and name.
* We needed to create 2 CTEs.
* The first is the profit by month.
* I created a Case using the released date since we don’t have a column for month the movie was released and the profit.
* I then created the second CTE to find the max profit from each month
* We then join both CTEs on the conditions that the months are the same and that the profit equals the max profit
* Lastly, we order the results by the max profit descending to get the most profitable at the top.

Query:

A screen shot of a computer

AI-generated content may be incorrect.

Result:

A screenshot of a computer

AI-generated content may be incorrect.

From the results above we can determine that the most profitable movie was from June with was E.T. with $782,410,554. In second is Ghost with a profit of $483,703,557 in July. And, in 3rd we had Pretty Woman with a profit of $449,406,268 on March.

1. Most profitable movie depending on the region where movie is from

* This one is very similar to the last question except we are grouping countries into regions.
* First, we need to retrieve the region, profit, and name.
* We need to create 2 CTEs, one for the region’s profit and the second is finding the region’s max profit.
* For the first CTE we create a case to make a column to determine the region of the film, name, and the profit of the film,
* The second CTE find the max profitable film from each region using the previous CTE.
* Now on the main query we join both CTEs on the condition that the region is equal and the profit is equal to the max profit.

Query:

A computer screen shot of a program

AI-generated content may be incorrect.

Result:

A screenshot of a computer

AI-generated content may be incorrect.

From the results above I can determine that in North America the most profitable film was E.T. with a profit of $782,410,554. In second is the Oceania region with film Crocodile Dundee with a profit of $319,403,506. In third it is Europe with the film License to Kill with a profit of $124,167,015.

**Conclusion**

To end this project, I was excited to use CTEs and get some use of them to understand more on how to use them. I am excited to know about them and how much more complex they can be. As well as CTEs joins where also something I wanted to get more into and this project with the CTEs helped. The data was also interesting to find out about and how dominant the U.S. film market was back in the 80s and 90s. There are many classic and iconic films I thought to be at the top but weren’t. I hope to make more projects like this and find more interesting answers to from questions created by data.

Thank you.